Nonparametric two-sample tests of high dimensional mean vectors via random integration

Yunlu Jiang, Xueqin Wang, Canhong Wen, Yukang Jiang, and Heping Zhang

Jinan University

University of Science and Technology of China

Sun Yat-Sen University

Yale University

.

[1]Yunlu Jiang (tjiangyl@jnu.edu.cn) is Associate Professor, Department of Statistics, College of Economics, Jinan University, Guangzhou, GD 510632, China. Xueqin Wang (wangxq20@ustc.edu.cn) is Professor, Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, AH 230026, China. Canhong Wen (wench@ustc.edu.cn) is Associate Professor, Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, AH 230026 China. Yukang Jiang (jiangyk3@mail2.sysu.edu.cn) is Ph.D. student, Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, Guangzhou, GD 510275, China. Heping Zhang (heping.zhang@yale.edu) is Susan Dwight Bliss Professor, Department of Biostatistics Yale University School of Public Health, New Haven, CT 06520-8034, U.S.A. Jiang's research is partially supported by NSFC(12171203) and the Natural Science Foundation of Guangdong (2019A1515011830, 2022A1515010045). Wang's research is partially supported by NSFC(72171216, 71991474), the International Science & Technology cooperation program of Guangdong, China(2016B050502007), the Key Research and Development Program of Guangdong, China(2019B020228001), and Science and Technology Program of Guangzhou, China(202002030129). Wen's research is partially supported by National Science Foundation of China (12171449) and Natural Science Foundation of Anhui Province (BJ2040170017). Zhang's research is partially supported by the U.S. National Institutes of Health (R01HG010171 and R01MH116527) and NSF (DMS-2112711). The authors report there are no competing interests to declare.

Nonparametric two-sample tests of high dimensional mean vectors via random integration

Abstract

Testing the equality of the means in two samples is a fundamental statistical inferential problem. Most of the existing methods are based on the sum-of-squares or supremum statistics. They are possibly powerful in some situations, but not in others, and they do not work in a unified way. Using random integration of the difference, we develop a framework that includes and extends many existing methods, especially in high-dimensional settings, without restricting the same covariance matrices or sparsity. Under a general multivariate model, we can derive the asymptotic properties of the proposed test statistic without specifying a relationship between the data dimension and sample size explicitly. Specifically, the new framework allows us to better understand the test's properties and select a powerful procedure accordingly. For example, we prove that our proposed test can achieve the power of 1 when nonzero signals in the true mean differences are weakly dense with nearly the same sign. In addition, we delineate the conditions under which the asymptotic relative Pitman efficiency of our proposed test to its competitor is greater than or equal to 1. Extensive numerical studies and a real data example demonstrate the potential of our proposed test.

Keywords: Nonparametric two-sample test; High-dimensional mean; Random integration of the difference

1. INTRODUCTION

In many applications, high-dimensional data, whose dimension is much larger than the sample size, are commonly available. Examples include diffusion tensor imaging (Le Bihan et al., 2001), finance (Lam and Yao, 2012), gene expression (Pan et al., 2018), and risk management (Bollerslev et al., 2019). Testing the equality of two high-dimensional mean vectors is a basic problem. For example, Chen and Qin (2010) and Zhang et al. (2020) studied differential gene expression in various molecules and tissues. Ayyala et al. (2015) detected differentially methylated regions based on MethylCap-seq data.

We deal with the two-sample test for equality of high-dimensional mean vectors. Given $\{\mathbf{X}_1, \cdots, \mathbf{X}_m\}$ and $\{\mathbf{Y}_1, \cdots, \mathbf{Y}_n\}$ are independent identically distributed random samples drawn from p-dimensional random variables $\mathbf{X}$ and $\mathbf{Y}$ having $p \times 1$ mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, we want to test:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad vs \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \tag{1.1}$$

where their covariance matrices Σ_1 and Σ_2 are unknown.

For testing (1.1), some existing methods assume that $\Sigma_1 = \Sigma_2$ (Bai and Saranadasa, 1996; Wu et al., 2006; Srivastava and Du, 2008; Li et al., 2020). This assumption is complicated to be validated for high-dimensional data. Motivated by the idea from Bai and Saranadasa (1996), Chen and Qin (2010) introduced a test statistic by removing the cross-product terms in $||\bar{\mathbf{X}}-\bar{\mathbf{Y}}||_2^2$, and showed that their proposed test can work with unequal covariance matrices. Also, Srivastava et al. (2013) extended the results of Srivastava and Du (2008) to unequal covariance matrices by replacing the sample covariance with the diagonal matrix of the sample covariance. Gregory et al. (2015) proposed using an average of the squared univariate two-sample t-statistics over p components as the test statistic. Wang and Xu (2022) proposed an approximate randomization test procedure based on the test statistic of Chen and Qin (2010). These test statistics use weighted L_2-norm between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, which are called as the sum-of-squares type statistics. It is known that sum-of-squares-based tests can have

good power against the "dense" alternatives, but otherwise, they may suffer from power loss (Cai et al., 2014; Xu et al., 2016).

Many methods deal with the sparsity for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ under the alternative hypothesis. Cai et al. (2014) used extreme value theory, which was heavily reliant on the dependence structure among various components of the p-dimensional random vectors. If there is a high degree of dependence, such as genetic data, their method based on extreme value theory fails, and the associated empirical size is distorted. Chang et al. (2017) calculated the critical value using the Gaussian approximation technique, which allowed for arbitrary dependency among different components of the p-dimensional random vectors. Besides, Chang et al. (2017) proposed a screening-based procedure for improving the supremum-type statistic's power. Cai et al. (2014) concluded that the supremum-based tests were powerful when the true mean differences were sparse in the sense that there were only a few but significant nonzero componentwise differences. However, such tests may not work well under a non-sparse alternative (Xu et al., 2016).

In practice, the true alternative hypothesis is unknown, so it can not help us choose a powerful test. Fortunately, powerful methods exist for both "dense" alternatives and sparse alternatives in the high-dimensional setting. Xu et al. (2016) proposed an adaptive testing procedure by combining information across a class of sum-of-powers tests. Chen et al. (2019) introduced an L_2-type test by either thresholding, which removed the non-signal bearing dimensions or transforming the data via the precision matrix for signal enhancement. Zhang et al. (2020) proposed an L^2-norm-based test through the Welch-Satterthwaite χ^2-approximation to deal with the non-normality of the null distribution. However, their test again assumes an equal covariance matrix between the two populations. In contrast, Xue and Yao (2020) proposed a distribution and correlation free two-sample mean test. Yu et al. (2022) proposed a power-enhanced high-dimensional mean test.

We categorize the methods mentioned above in Table 1 according to a) whether a test needs distribution assumptions; b) whether a test needs assumptions on the common covari-

ance matrix; c) whether a test needs a sparsity assumption under the alternative hypothesis; and d) whether a test requires clear conditions in the relationship between the data dimension p and the sample size n. Note that most of the methods are either sum-of-squares or supremum-based. Xu et al. (2016) pointed out that such tests were not powerful if nonzero signals in the true mean differences were weakly dense with nearly the same sign or there were more dense or only weakly dense nonzero signals, but did not offer a solution.

This paper establishes a unified framework by using random integration (Jiang et al., 2022) of the difference (RID) technique for two-sample tests of high-dimensional mean vectors. This technique uses the difference in the p-dimensional independent density-weighted function with the finite mean and variance. Many existing tests, such as the weighted L_2-norm-based test, the supremum-type tests, and a burden test through $\sum_{i=1}^{p}(\bar{X}^{(i)}-\bar{Y}^{(i)})$ (Pan and Shen, 2011; Lee et al., 2012), are special cases of our unified framework. Furthermore, our framework (RID) has the following advantages:

- It is nonparametric and can operate without assuming $\Sigma_1 = \Sigma_2$.
- It does not require a direct relationship between the data dimension and sample size, and nor the sparsity assumption under the alternative hypothesis.
- The asymptotic relative Pitman efficiency of our proposed RID test compared to the test (CQ) proposed by Chen and Qin (2010) is greater than or equal to 1 under some conditions.
- It leads to a distinctly powerful test when nonzero signals are weakly dense with nearly the same sign or when nonzero signals are "dense" under the alternative hypothesis. Hence, we solve the problem raised by Xu et al. (2016).

The rest of the paper is organized as follows. Section 2 introduces our test statistic via the random integration of the difference technique and establishes the asymptotic properties. In Section 3, simulation studies are conducted to evaluate the finite sample performance of

Table 1: The comparison of two-sample tests of high-dimensional mean vectors.

Methods	Distribution assumptions	Common covariance matrix	Sparsity	Assumptions between p and n	Power problem as (Xu et al., 2016) stated
BS (Bai and Saranadasa, 1996)	Yes	Yes		Yes	Yes
CQ (Chen and Qin, 2010)					Yes
Cai (Cai et al., 2014)			Yes	Yes	Yes
aSPU (Xu et al., 2016)				Yes	Yes
Mult1 (Chen et al., 2019)				Yes	Yes
L2 (Zhang et al., 2020)		Yes			Yes
DCF (Xue and Yao, 2020)				Yes	Yes

the proposed test. In Section 4, a real dataset is analyzed to compare the proposed test with some existing methods. We conclude with some remarks in Section 5. All technical details and some additional simulation results are provided as supplementary materials.

2. METHODOLOGY AND MAIN RESULTS

Note that

$$\begin{aligned}\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 &\Leftrightarrow E\mathbf{X} = E\mathbf{Y} \Leftrightarrow \boldsymbol{\delta}^\top E\mathbf{X} = \boldsymbol{\delta}^\top E\mathbf{Y}, \text{for any } \boldsymbol{\delta} \in \mathbb{R}^p \\ &\Leftrightarrow E\left[\boldsymbol{\delta}^\top(\mathbf{X}-\mathbf{Y})\right] = 0, \text{for any } \boldsymbol{\delta} \in \mathbb{R}^p.\end{aligned}$$

Therefore, testing whether $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ amounts to testing whether

$$\mathrm{RID}_w(\mathbf{X},\mathbf{Y}) \triangleq \int E^2\left[\boldsymbol{\delta}^\top(\mathbf{X}-\mathbf{Y})\right] w(\boldsymbol{\delta})d\boldsymbol{\delta} = 0, \tag{2.1}$$

where $w(\boldsymbol{\delta})$ is any positive weight.

It is important to clarify that for convenience, we use $\mathrm{RID}_w(\mathbf{X},\mathbf{Y})$ to indicate its dependence on the distributions of $\mathbf{X}$ and $\mathbf{Y}$, not the random variables $\mathbf{X}$ and $\mathbf{Y}$ per se. We obtain that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ if and only if $\mathrm{RID}_w(\mathbf{X},\mathbf{Y}) = 0$ by equation (2.1). Theorem 1 is a critical result that provides an explicit derivation for evaluating $\mathrm{RID}_w(\mathbf{X},\mathbf{Y})$ with a suitable w.

Theorem 1 *If $w(\boldsymbol{\delta}) = \prod_{i=1}^p w_i(\delta_i)$ and $w_i(\cdot)$ is a density function with a mean α_i and variance β_i^2 for $i = 1, \cdots, p$, then*

$$\begin{aligned}RID_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Y}) &\triangleq RID_w(\mathbf{X},\mathbf{Y}) \\ &= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top B(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \left[(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\mathbf{a}\right]^2,\end{aligned} \tag{2.2}$$

and $RID_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Y}) \geq 0$ with the equality holds if and only if $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, where $\boldsymbol{\theta} = (\alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_p)^\top$, $\mathbf{a} = (\alpha_1, \alpha_2, \cdots, \alpha_p)^\top$, and

$$B = \begin{pmatrix} \beta_1^2 & 0 & \cdots & 0 \\ 0 & \beta_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_p^2 \end{pmatrix}.$$

Remark 1 *Using Theorem 1, we can derive an explicit form of* $RID_w(\mathbf{X}, \mathbf{Y})$*, such as when* $\boldsymbol{\delta}$ *follows a density function with independent components. With different choices of the parameters* $\{\alpha_i, i = 1, \cdots, p\}$ *and* $\{\beta_i, i = 1, \cdots, p\}$*,* $RID_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})$ *can give rise to existing tests. Thus,* $RID_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})$ *provides a unified framework.*

1. *When* $\alpha_i = 0$ *and* $\beta_i \neq 0$ *for all* i*,* $RID_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})$ *yields a weighted* L_2*-norm-based test that is designed to be powerful for the "dense" alternatives (Chen and Qin, 2010).*

2. *When* $\alpha_1 = \cdots = \alpha_p \neq 0$ *and* $\beta_i = 0$ *for all* i*,* $RID_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})$ *results in a burden test, which is widely used in genome wide association study of rare variants (Pan and Shen, 2011; Lee et al., 2012).*

3. *When* $\alpha_i = 0$ *for all* i*,* $\beta_j \neq 0$ *for* $j = j_0$*, and otherwise* $\beta_j = 0$*, where* $j_0 = \arg\max_{1 \leq j \leq p} (\mu_{j1} - \mu_{j2})^2$*,* $RID_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})$ *can lead to the supremum-type tests using the* L_∞*-norm of the mean differences. In practice,* j_0 *is not known a priori and can be estimated by* $\hat{j}_0 = \arg\max_{1 \leq j \leq p} \left(\bar{\mathbf{X}}^{(j)} - \bar{\mathbf{Y}}^{(j)}\right)^2$*, where* $\bar{\mathbf{X}}^{(j)}$ *and* $\bar{\mathbf{Y}}^{(j)}$ *are the* j*-th component of sample mean vectors* $\bar{\mathbf{X}}$ *and* $\bar{\mathbf{Y}}$ *for* $\boldsymbol{\mu}_1$ *and* $\boldsymbol{\mu}_2$*, respectively. These tests are powerful against the "sparse" alternatives (Cai et al., 2014).*

4. *When* $\alpha_1 = \cdots = \alpha_p \neq 0$ *and* $\beta_i \neq 0$ *for all* i*,* $RID_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})$ *produces a hybrid of a weighted* L_2*-norm-based test and a burden test, and may retain the strengths of both tests with proper weights so that it is powerful whether there is a large proportion of small to moderate componentwise differences or nonzero signals are weakly dense with nearly the same sign (Chen and Qin, 2010; Xu et al., 2016).*

5. *When* $\beta_j \neq 0$ *for all* j*,* $\alpha_i > 0$ *if* $\mu_{i1} - \mu_{i2} > 0$*, and* $\alpha_i < 0$ *otherwise,* $RID_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})$ *induces a hybrid of a weighted* L_2*-norm-based test and a weighted* L_1*-norm-based test, which should be powerful for dense or only weakly dense nonzero signals (Chen and Qin, 2010; Xu et al., 2016).*

Denote $W_{\boldsymbol{\theta}} = B + \mathbf{a}\mathbf{a}^\top$. Then, we have

$$\mathrm{RID}_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y}) = ||W_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)||_2^2.$$

With observed samples $\{\mathbf{X}_1, \cdots, \mathbf{X}_m\}$ and $\{\mathbf{Y}_1, \cdots, \mathbf{Y}_n\}$, we define the test statistic as

$$\mathrm{RID}_{\boldsymbol{\theta},m,n} = \mathrm{RID}^1_{\boldsymbol{\theta},m} + \mathrm{RID}^2_{\boldsymbol{\theta},n} - 2\mathrm{RID}^3_{\boldsymbol{\theta},m,n},$$

where

$$\begin{aligned}
\mathrm{RID}^1_{\boldsymbol{\theta},m} &= \frac{1}{C^2_m} \sum_{1\le i<j\le m} \mathbf{X}_i^\top W_{\boldsymbol{\theta}} \mathbf{X}_j, \\
\mathrm{RID}^2_{\boldsymbol{\theta},n} &= \frac{1}{C^2_n} \sum_{1\le i<j\le n} \mathbf{Y}_i^\top W_{\boldsymbol{\theta}} \mathbf{Y}_j, \\
\mathrm{RID}^3_{\boldsymbol{\theta},m,n} &= \frac{1}{C^1_m C^1_n} \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{X}_i^\top W_{\boldsymbol{\theta}} \mathbf{Y}_j.
\end{aligned}$$

Obviously, $\mathrm{RID}_{\boldsymbol{\theta},m,n}$ is an unbiased estimator of $\mathrm{RID}_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})$.

2.1 Asymptotic properties

To establish the limiting distribution of $\mathrm{RID}_{\boldsymbol{\theta},m,n}$, we assume the following four conditions:

E1. There exist a $p\times k_1$ matrix Γ_1, a $p\times k_2$ matrix Γ_2, k_1-dimensional random vectors $\{\mathbf{Z}_{1i}\}_{i=1}^m$, and k_2-dimensional random vectors $\{\mathbf{Z}_{2j}\}_{j=1}^n$, such that $\mathbf{X}_i = \boldsymbol{\mu}_1 + \Gamma_1 \mathbf{Z}_{1i}$ for $i = 1, \cdots, m$, and $\mathbf{Y}_j = \boldsymbol{\mu}_2 + \Gamma_2 \mathbf{Z}_{2j}$ for $j = 1, \cdots, n$. And Γ_1, Γ_2, $\{\mathbf{Z}_{1i}\}_{i=1}^m$, and $\{\mathbf{Z}_{2j}\}_{j=1}^n$ satisfy:

1 $\Gamma_1\Gamma_1^\top = \Sigma_1$, and $\Gamma_2\Gamma_2^\top = \Sigma_2$ with $\min\{k_1, k_2\} \ge p$.

2 $\{\mathbf{Z}_{1i}\}_{i=1}^m$ and $\{\mathbf{Z}_{2j}\}_{j=1}^n$ are i.i.d., respectively, with $E\mathbf{Z}_{1i} = \mathbf{0}$, $\mathrm{Var}(\mathbf{Z}_{1i}) = I_{k_1}$, and $E\mathbf{Z}_{2j} = \mathbf{0}$, $\mathrm{Var}(\mathbf{Z}_{2j}) = I_{k_2}$, where I_{k_1} and I_{k_2} are the $k_1 \times k_1$ and $k_2 \times k_2$ identity matrices, respectively.

3 $E(Z^4_{1jl}) = 3 + \Delta_1$ and $E(Z^4_{2jl}) = 3 + \Delta_2$ for some constants Δ_1 and Δ_2, where $Z_{\iota jl}$ is the l-th component of $\mathbf{Z}_{\iota j}$ with $\iota = 1$ or 2. Also, for a positive integer q and ς_l such that $\sum_{l=1}^q \varsigma_l \le 8$,

$$E(Z^{\varsigma_1}_{ijl_1} \cdots Z^{\varsigma_q}_{ijl_q}) = E(Z^{\varsigma_1}_{ijl_1}) \cdots E(Z^{\varsigma_q}_{ijl_q}) \tag{2.3}$$

whenever $l_1, l_2, \cdots, l_q$ are distinct indices.

E2. $m/(m+n) \to \tau \in (0,1)$ as $m, n \to \infty$.

E3. $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top W_{\boldsymbol{\theta}} \Sigma_i W_{\boldsymbol{\theta}} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o[(m+n)^{-1} tr\{(W_{\boldsymbol{\theta}}\Sigma_1 + W_{\boldsymbol{\theta}}\Sigma_2)^2\}]$ for $i = 1$ or 2.

E4. As $p \to \infty$, and for $s_1, s_2, s_3, s_4 \in \{1, 2\}$,

$$tr(W_{\boldsymbol{\theta}}\Sigma_{s_1} W_{\boldsymbol{\theta}}\Sigma_{s_2} W_{\boldsymbol{\theta}}\Sigma_{s_3} W_{\boldsymbol{\theta}}\Sigma_{s_4}) = o[tr^2\{(W_{\boldsymbol{\theta}}\Sigma_1 + W_{\boldsymbol{\theta}}\Sigma_2)^2\}]. \tag{2.4}$$

Remark 2 *Condition E1 gives a general multivariate model for high-dimensional data analysis, which includes the Gaussian family and members of the elliptically contoured distributions among many others (Bai and Saranadasa, 1996; Chen and Qin, 2010; Zhang et al., 2020). As said in Chen and Qin (2010),* $\min\{k_1, k_2\} \geq p$ *indicates that the rank and eigenvalues of* Σ_1 *or* Σ_2 *are not affected by the transformation. Condition (2.3) means that each* $\mathbf{Z}_{ij}$ *has a kind of pseudo-independence, which is a relaxed independence relation that allows some margin over probabilities (Kim and Lesser, 2008). Clearly, if* $\mathbf{Z}_{ij}$ *has independent components, then (2.3) is true.*

Condition E2 is a standard regularity assumption in two-sample problems, which guarantees that m *and* n *go to infinity proportionally.*

Condition E3 is satisfied under H_0*, and enables the variance of* $RID_{\boldsymbol{\theta},m,n}$ *to be asymptotically characterized by* $\sigma^2_{m,n}$ *given in the following Theorem 2. Similar to Chen and Qin (2010), if all of the eigenvalues of* $W_{\boldsymbol{\theta}}$*,* Σ_1 *and* Σ_2 *are bounded above from infinity and below away from zero and* $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = (\omega, \cdots, \omega)^\top$*, then condition E3 implies* $\omega = o((n+m)^{-1/2})$*, which is also studied in Xu et al. (2016), and is a smaller order than the local alternative hypotheses with the form* $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\nu}(n+m)^{-1/2}$ *for a nonzero constant vector* $\boldsymbol{\nu}$ *and the fixed* p *setting. Therefore, condition E3 can be viewed as a high-dimensional version of the local alternative hypotheses.*

Condition E4 is typical to obtain a normal limit for the leading terms using the martingale central limit theorem (Hall, 1984), and similar conditions can be found in Chen et al. (2010) and Li and Chen (2012) for proving the asymptotic distribution in high-dimensional hypothesis-testing problems. In addition, condition E4 is also useful for our proposed test in the high-dimensional case, although an explicit relationship between p *and* m, n *is not*

required. Furthermore, if Σ_1 *and* Σ_2 *are close to identity matrix, then one must rule out the case that* $\beta_j = 0$ *for all* j*; otherwise condition E4 is violated. Therefore, other tests would be preferrable in this case. To gain insight into condition E4, let us assume* $\Sigma_1 = \Sigma_2 = \Sigma$, $\alpha_1 = \cdots = \alpha_p = \alpha \neq 0$, $\beta_1 = \cdots = \beta_p = \beta \neq 0$, $r = \alpha/\beta$, *and* $\lambda_1 \leq \lambda_2 \cdots \leq \lambda_p$ *and* $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_p$ *are eigenvalues of* $W_{\boldsymbol{\theta}}$ *and* Σ*, respectively. Then, (2.4) becomes*

$$tr\{(W_{\boldsymbol{\theta}}\Sigma)^4\} = o[tr^2\{(W_{\boldsymbol{\theta}}\Sigma)^2\}]. \tag{2.5}$$

By some algebraic calculations, we have $\lambda_1 = \cdots = \lambda_{p-1} = \beta^2$, *and* $\lambda_p = \beta^2 + p\alpha^2$, *and*

$$\begin{aligned} tr\{(W_{\boldsymbol{\theta}}\Sigma)^4\} &\leq \sum_{i=1}^{p}(\lambda_i\gamma_i)^4 = \lambda_1^4 tr(\Sigma^4) + (\lambda_p^4 - \lambda_1^4)\gamma_p^4, \\ tr\{(W_{\boldsymbol{\theta}}\Sigma)^2\} &\geq \sum_{i=1}^{p}(\lambda_i\gamma_{p-i+1})^2 \geq \lambda_1^2 tr(\Sigma^2). \end{aligned}$$

Therefore,

$$\frac{tr\{(W_{\boldsymbol{\theta}}\Sigma)^4\}}{tr^2\{(W_{\boldsymbol{\theta}}\Sigma)^2\}} \leq \frac{\lambda_1^4 tr(\Sigma^4) + (\lambda_p^4 - \lambda_1^4)\gamma_p^4}{\lambda_1^4 tr^2(\Sigma^2)} = \frac{tr(\Sigma^4)}{tr^2(\Sigma^2)} + \frac{[(1+pr^2)^4 - 1]\gamma_p^4}{tr^2(\Sigma^2)}.$$

Thus, if $tr(\Sigma^4) = o(tr^2(\Sigma^2))$, $pr^2 = O(p^{1/4})$, *and* $p^{1/2}\gamma_p^2 = o(tr(\Sigma^2))$, *then (2.5) is true. In fact,* $tr(\Sigma^4) = o(tr^2(\Sigma^2))$ *is used in Chen et al. (2010) and Li and Chen (2012). If all the eigenvalues of* Σ *are bounded away from zero and infinity,* $tr(\Sigma^4) = o(tr^2(\Sigma^2))$ *is trivially true. Meanwhile, some of the commonly encountered covariance structures satisfy* $tr(\Sigma^4) = o(tr^2(\Sigma^2))$, *see Chen et al. (2010). In addition,* $p^{1/2}\gamma_p^2 = o(tr(\Sigma^2))$ *is also assumed in the literature, e.g., Wang et al. (2015).*

Theorem 2 *Under Conditions E1-E4, as* $p, m, n \to \infty$, *we have*

$$\frac{RID_{\boldsymbol{\theta},m,n} - RID_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})}{\sigma_{m,n}} \xrightarrow{\mathscr{D}} \mathcal{N}(0, 1),$$

where

$$\sigma_{m,n}^2 = \frac{tr\{(W_{\boldsymbol{\theta}}\Sigma_1)^2\}}{C_m^2} + \frac{tr\{(W_{\boldsymbol{\theta}}\Sigma_2)^2\}}{C_n^2} + \frac{4tr(W_{\boldsymbol{\theta}}\Sigma_1 W_{\boldsymbol{\theta}}\Sigma_2)}{C_m^1 C_n^1}.$$

Under H_0, we can obtain $\mathrm{RID}_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y}) = 0$. Therefore, we have the following Corollary 1.

Corollary 1 *Under Conditions E1-E4 and* $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, *as* $p, m, n \to \infty$, *we have*

$$RID_{\boldsymbol{\theta},m,n}/\sigma_{m,n} \xrightarrow{\mathscr{D}} \mathcal{N}(0,1).$$

To formulate a test procedure, we need to estimate $\sigma_{m,n}$. Similar to Chen and Qin (2010), we propose the following estimators of $tr\{(W_{\boldsymbol{\theta}}\Sigma_1)^2\}$, $tr\{(W_{\boldsymbol{\theta}}\Sigma_2)^2\}$, and $tr(W_{\boldsymbol{\theta}}\Sigma_1 W_{\boldsymbol{\theta}}\Sigma_2)$. Denote

$$\begin{aligned}
tr\{\widehat{(W_{\boldsymbol{\theta}}\Sigma_1)^2}\} &= \frac{1}{2C_m^2} tr\left\{\sum_{i\neq j} W_{\boldsymbol{\theta}}^{1/2}(\mathbf{X}_i - \bar{\mathbf{X}}_{(i,j)})\mathbf{X}_i^\top W_{\boldsymbol{\theta}}(\mathbf{X}_j - \bar{\mathbf{X}}_{(i,j)})\mathbf{X}_j^\top W_{\boldsymbol{\theta}}^{1/2}\right\},\\
tr\{\widehat{(W_{\boldsymbol{\theta}}\Sigma_2)^2}\} &= \frac{1}{2C_n^2} tr\left\{\sum_{i\neq j} W_{\boldsymbol{\theta}}^{1/2}(\mathbf{Y}_i - \bar{\mathbf{Y}}_{(i,j)})\mathbf{Y}_i^\top W_{\boldsymbol{\theta}}(\mathbf{Y}_j - \bar{\mathbf{Y}}_{(i,j)})\mathbf{Y}_j^\top W_{\boldsymbol{\theta}}^{1/2}\right\},\\
tr(\widehat{W_{\boldsymbol{\theta}}\Sigma_1 W_{\boldsymbol{\theta}}\Sigma_2}) &= \frac{1}{C_m^1 C_n^1} tr\left\{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{\boldsymbol{\theta}}^{1/2}(\mathbf{X}_i - \bar{\mathbf{X}}_{(i)})\mathbf{X}_i^\top W_{\boldsymbol{\theta}}(\mathbf{Y}_j - \bar{\mathbf{Y}}_{(j)})\mathbf{Y}_j^\top W_{\boldsymbol{\theta}}^{1/2}\right\},
\end{aligned}$$

where $\bar{\mathbf{X}}_{(i,j)}$ and $\bar{\mathbf{Y}}_{(i,j)}$ are the sample mean after excluding $\mathbf{X}_i, \mathbf{X}_j$ and $\mathbf{Y}_i, \mathbf{Y}_j$, respectively, and $\bar{\mathbf{X}}_{(i)}$ and $\bar{\mathbf{Y}}_{(j)}$ are the sample mean after excluding $\mathbf{X}_i$ and $\mathbf{Y}_j$, respectively. Therefore, we can obtain an estimator of $\sigma_{m,n}^2$.

$$\hat{\sigma}_{m,n}^2 = \frac{tr\{\widehat{(W_{\boldsymbol{\theta}}\Sigma_1)^2}\}}{C_m^2} + \frac{tr\{\widehat{(W_{\boldsymbol{\theta}}\Sigma_2)^2}\}}{C_n^2} + \frac{4tr(\widehat{W_{\boldsymbol{\theta}}\Sigma_1 W_{\boldsymbol{\theta}}\Sigma_2})}{C_m^1 C_n^1}.$$

Furthermore, we can obtain the following Theorem 3.

Theorem 3 *Under Conditions E1-E4 and* $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, *as* $p, m, n \to \infty$, *we have*

$$RID_{\boldsymbol{\theta},m,n}/\hat{\sigma}_{m,n} \xrightarrow{\mathscr{D}} \mathcal{N}(0,1).$$

According to Theorem 3, the proposed test with a nominal ϑ level of significance rejects H_0 if $\text{RID}_{\boldsymbol{\theta},m,n} \geq \hat{\sigma}_{m,n} z_\vartheta$, where z_ϑ is the upper-ϑ quantile of $\mathcal{N}(0,1)$.

2.2 Power of the proposed RID test

In this subsection, we investigate the power of the proposed RID test. Denote

$$\mathscr{P}_{\boldsymbol{\theta}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2) = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top W_{\boldsymbol{\theta}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\sqrt{2tr(\Sigma_{\boldsymbol{\theta}}^2(\tau))}},$$

where $\Sigma_{\boldsymbol{\theta}}(\tau) = W_{\boldsymbol{\theta}}\{(1-\tau)\Sigma_1 + \tau\Sigma_2\}$. Then, we can obtain the following Theorem 4.

Theorem 4 *Under Conditions E1-E4, and* $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, *the power of our proposed RID test is given by*

$$\begin{aligned}&\lim_{m,n,p\to\infty} P\left(RID_{\boldsymbol{\theta},m,n} \geq \hat{\sigma}_{m,n} z_\vartheta\right) \\ &= \lim_{m,n,p\to\infty} \Phi\left\{-z_\vartheta + (m+n)\tau(1-\tau)\mathscr{P}_{\boldsymbol{\theta}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)\right\},\end{aligned}$$

where $\Phi(\cdot)$ *is the cumulative distribution function of the standard normal random variable.*

Theorem 4 provides a general result on the power of the proposed RID test statistic. It reveals that as long as $(m+n)\mathscr{P}_{\boldsymbol{\theta}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$ diverges to the infinity, the power will converge to 1. Next, we investigate the power of the following two special cases for heterogeneous α_i and β_i. Let $\lambda_1 \leq \lambda_2 \cdots \leq \lambda_p$ and $\lambda_1^* \leq \lambda_2^* \cdots \leq \lambda_p^*$ be eigenvalues of $W_{\boldsymbol{\theta}}$ and $\widetilde{\Sigma}(\tau)$, respectively, where $\widetilde{\Sigma}(\tau) = (1-\tau)\Sigma_1 + \tau\Sigma_2$. For the sake of simplicity, let us assume that $\alpha_1 = \cdots = \alpha_p = \alpha$ and $\beta_1 \leq \cdots \leq \beta_p$. In this setting, we have $\lambda_1 = \beta_1^2, \cdots, \lambda_{p-1} = \beta_{p-1}^2$, and $\lambda_p = \beta_p^2 + \alpha^2\zeta_p$, where

$$\zeta_p = \frac{\beta_p^2}{\beta_1^2} + \frac{\beta_p^2}{\beta_2^2} + \cdots + \frac{\beta_p^2}{\beta_p^2}.$$

Case I: $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = (\omega, \cdots, \omega)^\top$ for $\omega \geq 0$. Then, we have

$$\mathscr{P}_{\boldsymbol{\theta}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2) \geq \frac{\alpha^2 p^2 \omega^2 + \omega^2 \sum_{i=1}^p \beta_i^2}{\sqrt{2(\lambda_p^*)^2\left(\sum_{i=1}^p \beta_i^4 + 2\beta_p^2\alpha^2\zeta_p + \alpha^4\zeta_p^2\right)}}.$$

Therefore, we have

$$\begin{aligned}&\lim_{m,n,p\to\infty} P\left(\mathrm{RID}_{\boldsymbol{\theta},m,n} \geq \hat{\sigma}_{m,n} z_\vartheta\right) \\ &\geq \lim_{m,n,p\to\infty} \Phi\left\{-z_\vartheta + \frac{(m+n)\tau(1-\tau)(\alpha^2 p^2\omega^2 + \omega^2\sum_{i=1}^p \beta_i^2)}{\sqrt{2(\lambda_p^*)^2\left(\sum_{i=1}^p \beta_i^4 + 2\beta_p^2\alpha^2\zeta_p + \alpha^4\zeta_p^2\right)}}\right\}.\end{aligned}$$

By the above inequality, we can obtain the following Corollary 2.

Corollary 2 *Assume Conditions E1-E4,* $\alpha_1 = \cdots = \alpha_p = \alpha$, $\beta_1 \leq \cdots \leq \beta_p$, $\alpha^2 = O(p^{-3/4})$, $0 < \min_{1\leq i\leq p}\{\beta_i\} \leq \max_{1\leq i\leq p}\{\beta_i\} < \infty$, *and* $\lambda_p^* = o\left((n+m)\omega^2 p^{1/8}\right)$. *Under* $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, *we have*

$$\lim_{m,n,p\to\infty} P\left(RID_{\boldsymbol{\theta},m,n} \geq \hat{\sigma}_{m,n} z_\vartheta\right) = 1.$$

Corollary 2 demonstrates that our proposed RID test is powerful when nonzero signals in the true mean differences are weakly dense with nearly the same sign. By contrast, (3.11) in Chen and Qin (2010) and (27) in Zhang et al. (2020) indicate that their tests have low power under H_1.

Case II: $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = (\overbrace{\varpi, \cdots, \varpi}^{p_1}, \overbrace{0, \cdots, 0}^{p-p_1})^\top$. Using a similar discussion to that of Corollary 2, we can obtain Corollary 3.

Corollary 3 *Assume Conditions E1-E4,* $\alpha_1 = \cdots = \alpha_p = \alpha$, $\beta_1 \leq \cdots \leq \beta_p$, $\alpha^2 = O(p^{-3/4})$, $0 < \min_{1\leq i\leq p}\{\beta_i\} \leq \max_{1\leq i\leq p}\{\beta_i\} < \infty$, $p_1 = p^e$ *with* $0 \leq e \leq 1$, *and* $\lambda_p^* = o\left((m+n)\omega^2 p^{\frac{3e}{2}-\frac{7}{8}}\right)$. *Under* $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, *we have*

$$\lim_{m,n,p\to\infty} P\left(RID_{\boldsymbol{\theta},m,n} \geq \hat{\sigma}_{m,n} z_\vartheta\right) = 1.$$

Corollary 3 indicates that if all the eigenvalues of $\widetilde{\Sigma}(\tau)$ are bounded away from 0 and $\omega^2 = O((m+n)^{-1})$, then $e > 7/12$. Therefore, the proposed RID test cannot cope with extremely sparse signals as did in Cai et al. (2014) and Chang et al. (2017) unless the sparse nonzero signals are extremely strong.

It follows from (2.2) that the first term in $\mathrm{RID}_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y})$ is a weighted L_2-norm between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Therefore, we will compare the asymptotic power of the our proposed RID test with the CQ test of Chen and Qin (2010). According to Theorem 4, and equation (3.11) in Chen and Qin (2010), the asymptotic power functions for the proposed RID test and CQ

test are defined as follows

$$\begin{aligned}\beta_{RID} &\triangleq \Phi\left\{-z_\vartheta + (m+n)\tau(1-\tau)\mathscr{P}_{\boldsymbol{\theta}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2,\Sigma_1,\Sigma_2)\right\},\\ \beta_{CQ} &\triangleq \Phi\left\{-z_\vartheta + \frac{(m+n)\tau(1-\tau)||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||^2}{\sqrt{2tr\{\widetilde{\Sigma}(\tau)^2\}}}\right\}.\end{aligned}$$

Then, the asymptotic relative Pitman efficiency of the proposed RID test versus the CQ test is given by

$$ARE(\beta_{RID},\beta_{CQ}) \triangleq \frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top W_{\boldsymbol{\theta}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\sqrt{tr\{\widetilde{\Sigma}(\tau)^2\}}}{||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||^2\sqrt{tr(\Sigma^2_{\boldsymbol{\theta}}(\tau))}}.$$

Furthermore, we can obtain the following Theorem 5.

Theorem 5 *Assume Conditions E1-E4,* $\alpha_1=\cdots=\alpha_p=\alpha$, $\beta_1=\cdots=\beta_p=\beta$, $r=\alpha/\beta$, *and* $\max\left\{pr^2(\lambda_p^*)^2, p^2r^4(\lambda_p^*)^2\right\} = o(tr\{\widetilde{\Sigma}(\tau)^2\})$. *Under* $H_1:\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$, *we have*

$$\lim_{m,n,p\to\infty} ARE(\beta_{RID},\beta_{CQ}) \geq 1.$$

Remark 3 *If* $pr^2 = O(p^{1/4})$ *and* $p^2r^4(\lambda_p^*)^2 = o(tr\{\widetilde{\Sigma}(\tau)^2\})$, *we have*

$$\max\left\{pr^2(\lambda_p^*)^2, p^2r^4(\lambda_p^*)^2\right\} = o(tr\{\widetilde{\Sigma}(\tau)^2\}).$$

The conditions, $pr^2 = O(p^{1/4})$ *and* $p^2r^4(\lambda_p^*)^2 = o(tr\{\widetilde{\Sigma}(\tau)^2\})$, *are implied by* $p^{1/2}(\lambda_p^*)^2 = o(tr\{\widetilde{\Sigma}(\tau)^2\})$. *The latter condition is used in the literature (e.g., Remark 3 in Wang et al. (2015)) and stronger than ours.*

3. SIMULATION STUDIES

Example 1. In this example, we investigate the numerical performance of the proposed method using Monte Carlo simulations in the presence of weak signals. We compare the RID with the adaptive sum-of-powers test proposed by Xu et al. (2016) (aSPU), the CQ test (Chen and Qin, 2010), the method without transformation proposed by Cai et al. (2014) (Cai), the non-studentized test with screening ($\Psi^f_{ns,\vartheta}$) (Chang et al., 2017), the multilevel thresholding

test without the data transformation proposed by Chen et al. (2019) (Mult1), an L^2-norm-based test proposed by Zhang et al. (2020) (L2), and the distribution and correlation free (DCF) two-sample mean test proposed by Xue and Yao (2020). The aSPU is implemented by the function apval_aSPU in the R package *highmean*. As suggested in Xu et al. (2016), to obtain aSPU, γ takes seven values, i.e., $\gamma \in \{1, 2, 3, 4, 5, 6, \infty\}$, and p-value is given by (1) in Xu et al. (2016).

For RID, according to Corollary 2, Corollary 3, and Theorem 5, we set $\alpha_1 = \cdots = \alpha_p = 2p^{-3/8}$ and $\beta_i = \sqrt{2}(p+i)/p, i = 1, \cdots, p$ for convenience. In this simulation, the nominal significance level is set to $\vartheta = 0.05$. For each setting, 1000 replications are simulated to calculate all empirical p-values and power levels.

The two random samples were generated according to the following model

$$\begin{aligned} \mathbf{X}_i &= \boldsymbol{\mu}_1 + \Sigma_1^{1/2}\mathbf{Z}_i \quad \text{for } i = 1, \cdots, m, \\ \mathbf{Y}_j &= \boldsymbol{\mu}_2 + \Sigma_2^{1/2}\mathbf{Z}_{m+j} \quad \text{for } j = 1, \cdots, n, \end{aligned}$$

where $\{\mathbf{Z}_i : i = 1, \cdots, m+n\}$ are independent p-dimensional random variables with i.i.d. coordinate $Z_{ik}, k = 1, \cdots, p$. We consider the following four distributions for Z_{ik}:

1. The standard normal $\mathcal{N}(0, 1)$;
2. The standardized t-distribution with degrees of freedom 5, i.e., $(5/3)^{-1/2}t(5)$;
3. The standardized chi-squared distribution with degrees of freedom 4, i.e., $8^{-1/2}\{\chi^2(4) - 4\}$;
4. The standardized Gamma distribution with $a = 4, b = 0.5$, i.e., $\Gamma(4, 0.5) - 2$.

We assigned $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$ under H_0 and under H_1, $\boldsymbol{\mu}_1 = \mathbf{0}$, and $\boldsymbol{\mu}_2$ had $[p^{1-\rho}]$ non-zero entries of equal value that were uniformly allocated among $\{1, \cdots, p\}$, where $[a]$ denotes the integer part of a. The values of the nonzero entries were $\sqrt{2r(1/m + 1/n)\log p}$, where

$r > 0$, and $\rho \in [0, 1]$ controls the signal sparsity. For the covariance matrix, we consider the following two scenarios:

Scenario 1: Unequal covariance matrices, $\Sigma_1 = \left(0.4^{|i-j|}\right)$ and $\Sigma_2 = 2\Sigma_1$ for $1 \leq i, j \leq p$.

Scenario 2: Common covariance matrices, $\Sigma_1 = \Sigma_2 = \Sigma$, where $\Sigma = \left(0.9^{|i-j|}\right)$ for $1 \leq i, j \leq p$.

In this simulation, we set the sample sizes to $(m, n) = (60, 80)$ and $(90, 120)$, respectively, and the dimension p to $200, 600, 1000$. We take $\rho = \{0.1, 0.2, 0.3, 0.4\}$, covering highly dense signals for an alternative hypothesis at $\rho = 0.1$, to moderately dense signals at $\rho = 0.2$ or $\rho = 0.3$, and finally to moderately sparse at $\rho = 0.4$. Meanwhile, we set the signal strength $r = \{0.02, 0.04, 0.06, 0.08\}$ which covers the same range as in Xu et al. (2016).

The empirical p-values are shown in Tables 2-3. From Tables 2-3, we find that the empirical p-values of all methods are controlled fairly well around 0.05 for all cases.

The power of RID is similar under the four different distributions, so we report the empirical power under only $\mathcal{N}(0, 1)$ in Figures 1-4. For the other three distributions, the results are presented in the supplementary material. From Figures 1-4, we have the following findings:

1. The proposed RID test has the greatest power when $\rho = 0.1$ or $\rho = 0.2$. This result is consistent with Corollary 2. Thus, the RID is powerful when nonzero signals of the difference between two mean vectors are weakly dense with nearly the same sign.

2. The empirical power of RID increases as the signal strength r increases.

3. RID's empirical power diminishes as ρ increases.

Table 2: Empirical sizes for **Scenario 1**

p	(m,n)	aSPU	CQ	Cai	Mult1	L2	DCF	$\Psi^f_{ns,\vartheta}$	RID
					$\mathcal{N}(0,1)$				
200	(60,80)	0.033	0.061	0.049	0.036	0.049	0.039	0.054	0.051
	(90,120)	0.035	0.049	0.054	0.026	0.040	0.038	0.058	0.052
600	(60,80)	0.047	0.056	0.072	0.026	0.038	0.048	0.056	0.049
	(90,120)	0.036	0.048	0.052	0.028	0.035	0.039	0.063	0.044
1000	(60,80)	0.044	0.045	0.064	0.021	0.017	0.032	0.047	0.055
	(90,120)	0.042	0.054	0.061	0.033	0.028	0.038	0.058	0.054
					$(5/3)^{-1/2}t(5)$				
200	(60,80)	0.033	0.060	0.037	0.030	0.042	0.022	0.043	0.045
	(90,120)	0.041	0.064	0.057	0.045	0.050	0.035	0.039	0.052
600	(60,80)	0.028	0.075	0.048	0.034	0.036	0.022	0.027	0.055
	(90,120)	0.035	0.049	0.054	0.029	0.030	0.030	0.032	0.052
1000	(60,80)	0.034	0.056	0.055	0.024	0.028	0.017	0.035	0.056
	(90,120)	0.031	0.044	0.043	0.016	0.025	0.023	0.035	0.053
					$8^{-1/2}\{\chi^2(4)-4\}$				
200	(60,80)	0.034	0.061	0.053	0.038	0.051	0.032	0.051	0.058
	(90,120)	0.033	0.055	0.046	0.031	0.043	0.036	0.050	0.051
600	(60,80)	0.038	0.050	0.060	0.022	0.027	0.030	0.042	0.054
	(90,120)	0.046	0.043	0.072	0.033	0.030	0.052	0.049	0.049
1000	(60,80)	0.030	0.054	0.077	0.020	0.020	0.024	0.043	0.057
	(90,120)	0.039	0.056	0.066	0.032	0.033	0.028	0.040	0.056
					$\Gamma(4,0.5)-2$				
200	(60,80)	0.042	0.050	0.053	0.032	0.033	0.037	0.061	0.052
	(90,120)	0.033	0.065	0.044	0.041	0.049	0.036	0.046	0.057
600	(60,80)	0.046	0.041	0.076	0.030	0.023	0.034	0.048	0.049
	(90,120)	0.051	0.053	0.061	0.034	0.039	0.033	0.043	0.048
1000	(60,80)	0.050	0.048	0.073	0.030	0.030	0.027	0.031	0.050
	(90,120)	0.033	0.065	0.068	0.037	0.041	0.036	0.045	0.047

Table 3: Empirical sizes for **Scenario 2**

p	(m,n)	aSPU	CQ	Cai	Mult1	L2	DCF	$\Psi^f_{ns,\vartheta}$	RID
					$\mathcal{N}(0,1)$				
200	(60,80)	0.049	0.078	0.034	0.046	0.063	0.051	0.058	0.048
	(90,120)	0.049	0.062	0.029	0.043	0.051	0.039	0.046	0.055
600	(60,80)	0.051	0.063	0.047	0.046	0.052	0.038	0.047	0.057
	(90,120)	0.055	0.057	0.032	0.041	0.051	0.042	0.051	0.047
1000	(60,80)	0.036	0.061	0.042	0.041	0.052	0.040	0.050	0.057
	(90,120)	0.042	0.053	0.044	0.037	0.046	0.038	0.051	0.055
					$(5/3)^{-1/2}t(5)$				
200	(60,80)	0.038	0.059	0.030	0.040	0.044	0.038	0.042	0.057
	(90,120)	0.042	0.060	0.025	0.044	0.048	0.045	0.052	0.059
600	(60,80)	0.033	0.051	0.029	0.033	0.039	0.032	0.039	0.055
	(90,120)	0.059	0.071	0.033	0.051	0.061	0.040	0.047	0.052
1000	(60,80)	0.034	0.051	0.035	0.036	0.046	0.023	0.028	0.049
	(90,120)	0.045	0.068	0.042	0.054	0.060	0.050	0.059	0.054
					$8^{-1/2}\{\chi^2(4)-4\}$				
200	(60,80)	0.049	0.079	0.043	0.052	0.063	0.057	0.062	0.055
	(90,120)	0.062	0.086	0.030	0.058	0.065	0.047	0.050	0.056
600	(60,80)	0.040	0.065	0.038	0.036	0.055	0.037	0.040	0.056
	(90,120)	0.039	0.055	0.037	0.032	0.046	0.041	0.044	0.055
1000	(60,80)	0.045	0.069	0.059	0.044	0.056	0.053	0.063	0.054
	(90,120)	0.042	0.060	0.047	0.040	0.050	0.048	0.055	0.054
					$\Gamma(4,0.5)-2$				
200	(60,80)	0.051	0.064	0.040	0.040	0.052	0.049	0.059	0.052
	(90,120)	0.048	0.066	0.024	0.040	0.056	0.045	0.051	0.057
600	(60,80)	0.054	0.065	0.042	0.037	0.051	0.046	0.055	0.050
	(90,120)	0.047	0.060	0.039	0.046	0.052	0.049	0.055	0.056
1000	(60,80)	0.036	0.065	0.044	0.040	0.051	0.044	0.055	0.049
	(90,120)	0.045	0.055	0.033	0.035	0.045	0.032	0.040	0.047

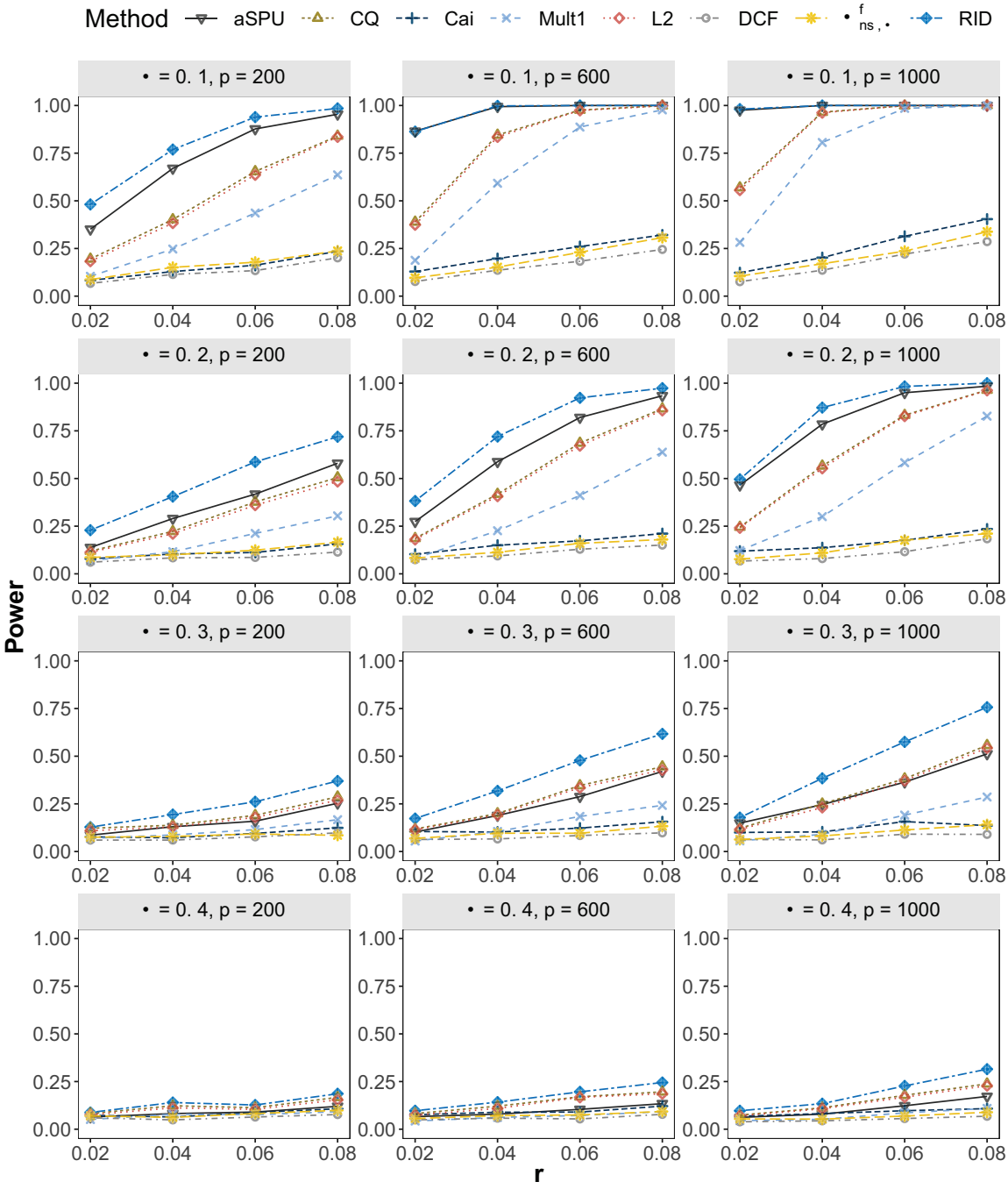


Figure 1: Empirical power when Z_{ij} follows $\mathcal{N}(0,1)$ and $m = 60, n = 80$ for **Scenario 1** under different signal levels of r and sparsity levels of ρ in **Example 1**.

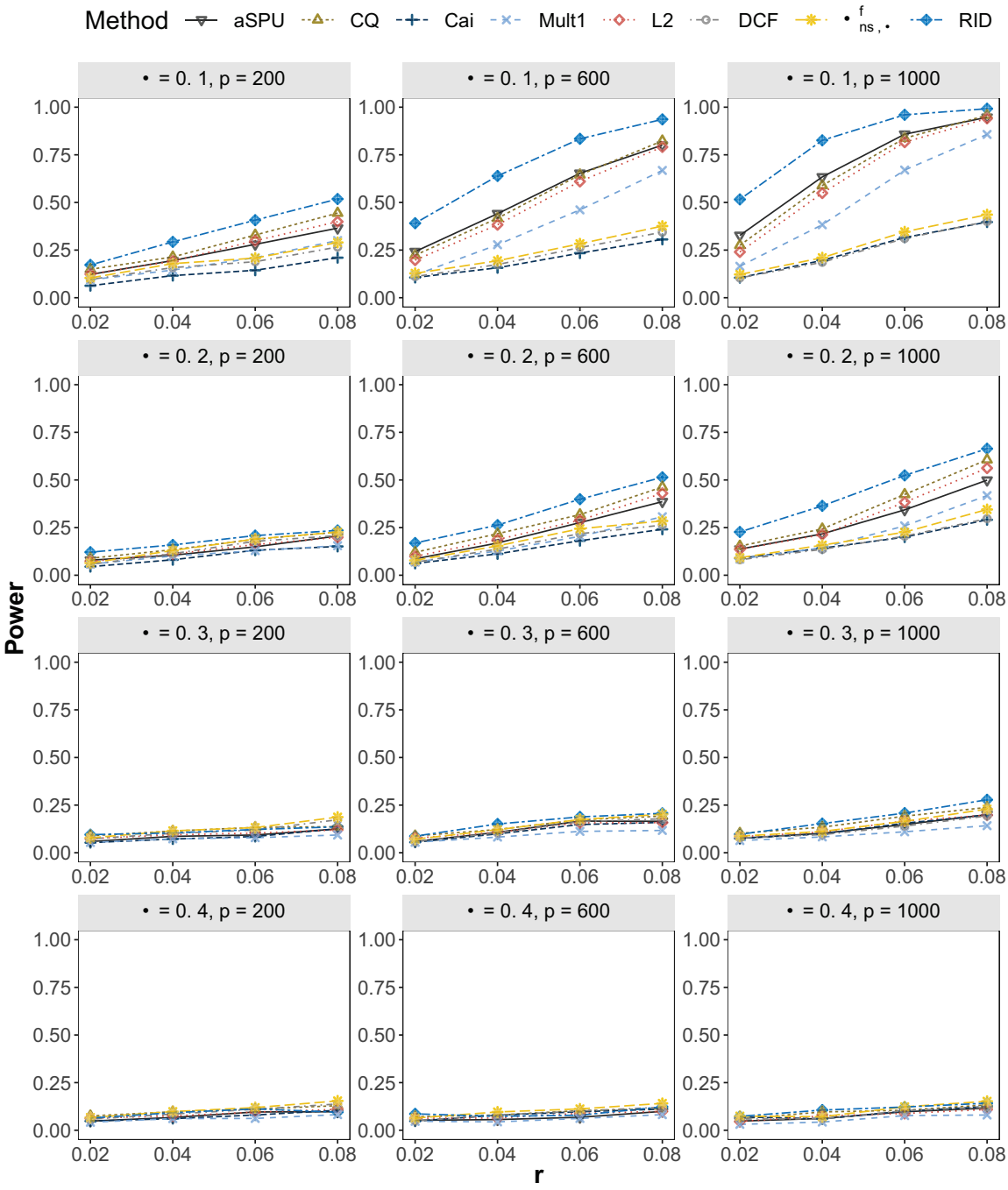


Figure 2: Empirical power when Z_{ij} follows $\mathcal{N}(0,1)$ and $m = 60, n = 80$ for **Scenario 2** under different signal levels of r and sparsity levels of ρ in **Example 1**.

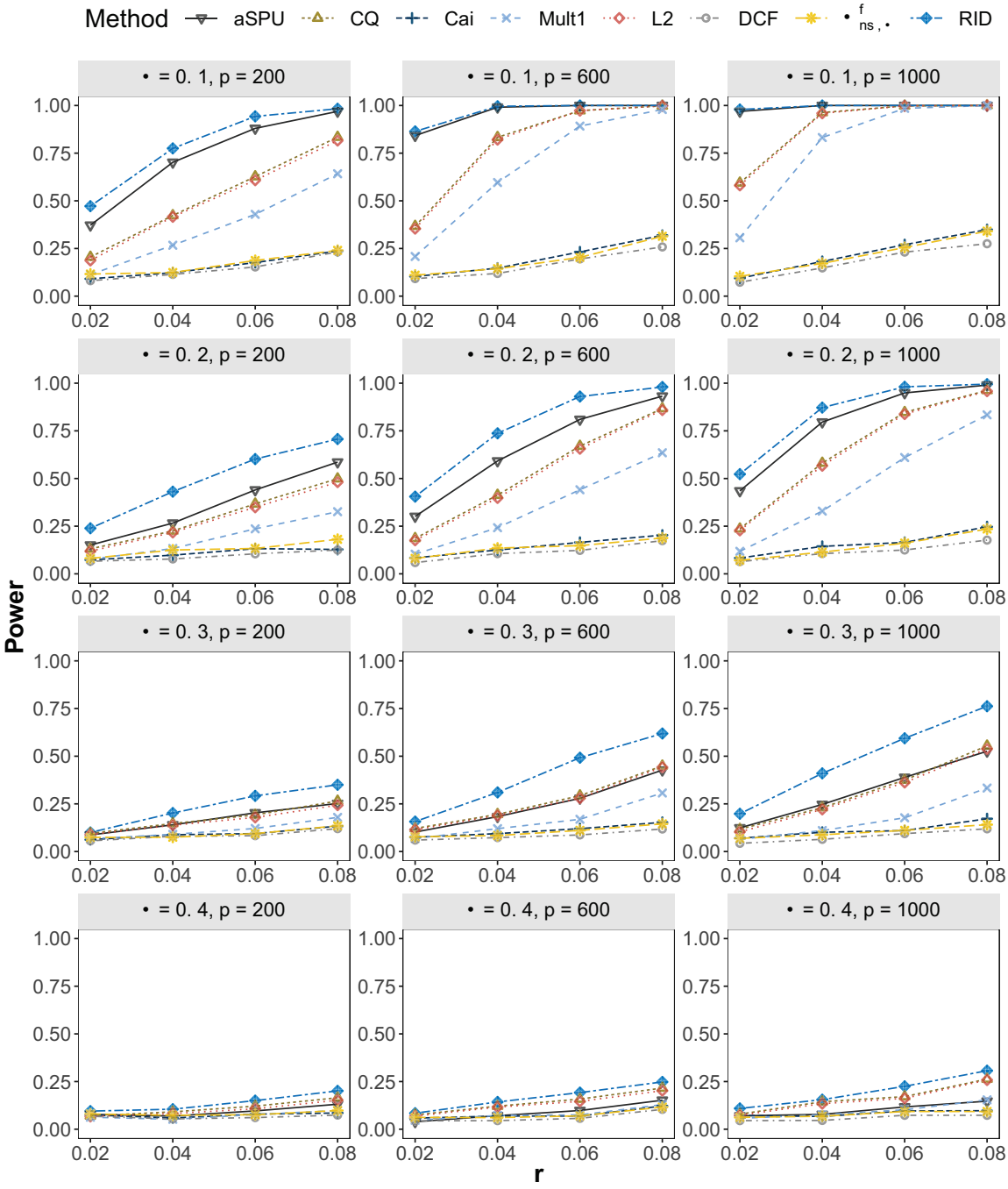


Figure 3: Empirical power when Z_{ij} follows $\mathcal{N}(0,1)$ and $m = 90, n = 120$ for **Scenario 1** under different signal levels of r and sparsity levels of ρ in **Example 1**.

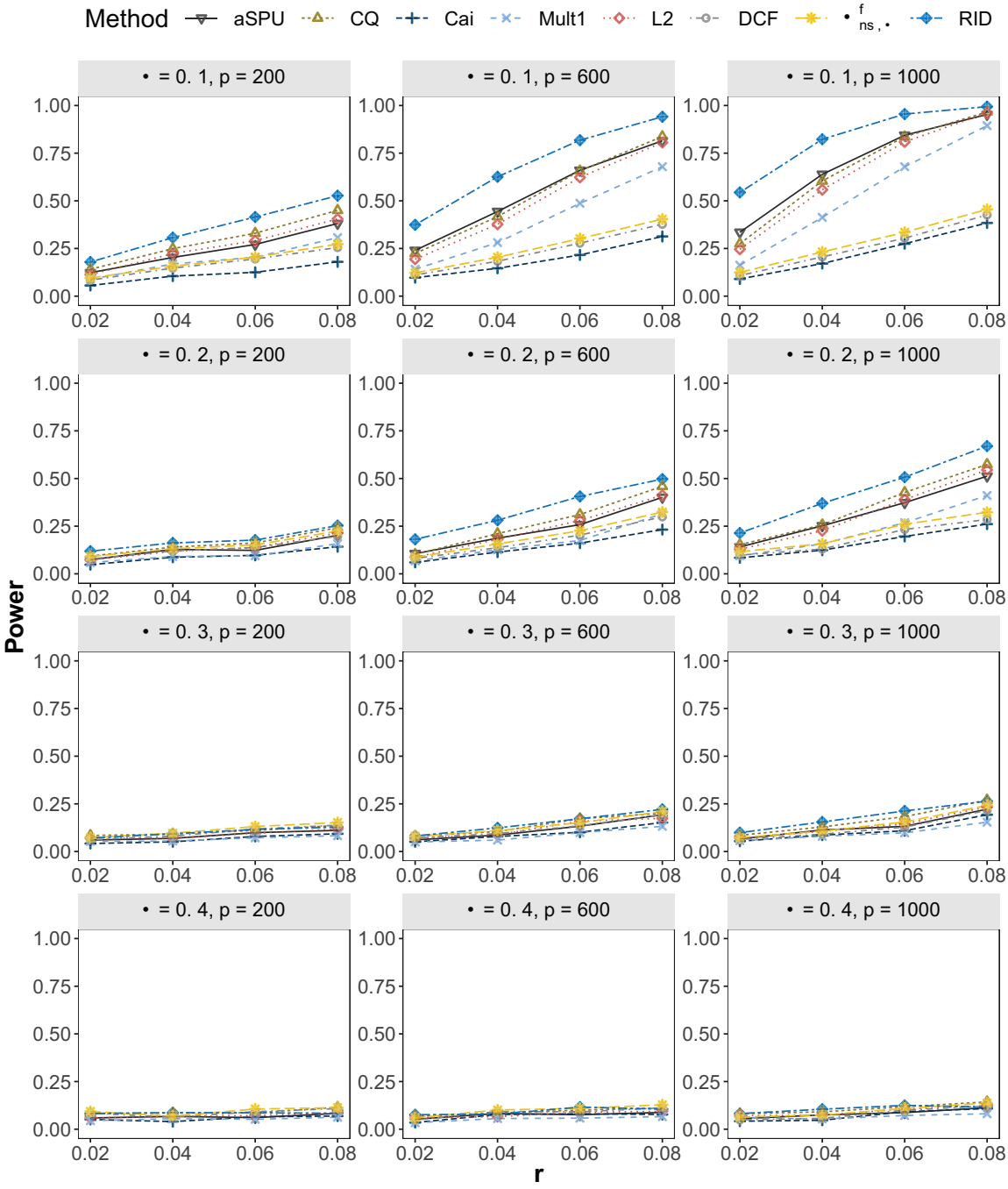


Figure 4: Empirical power when Z_{ij} follows $\mathcal{N}(0,1)$ and $m = 90, n = 120$ for **Scenario 2** under different signal levels of r and sparsity levels of ρ in **Example 1**.

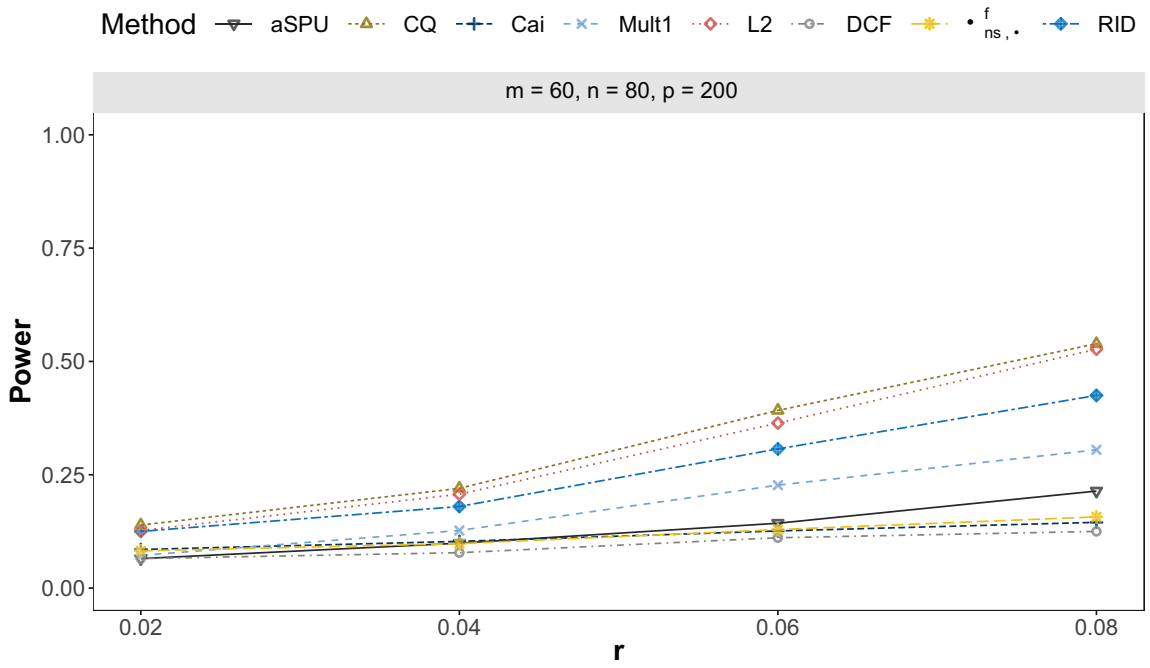


Figure 5: Empirical power when Z_{ij} follows $\mathcal{N}(0,1)$, $(m,n) = (60,80)$, $p = 200$ under different signal levels of r in **Example 2**.

Example 2. In this example, we evaluate the power performance of our proposed RID test for the weakly dense signals with varying signs. We use the same setup in **Example 1**, except that $\rho = 0$, $\Sigma_1 = \left(0.4^{|i-j|}\right)$ and $\Sigma_2 = 2\Sigma_1$ for $1 \leq i, j \leq p$, $(m,n) = (60,80)$, and $p = 200$. The values of the nonzero entries in $\boldsymbol{\mu}_2$ were uniformly drawn at random from the interval $\left[-\sqrt{2r(1/m+1/n)\log p}, \sqrt{2r(1/m+1/n)\log p}\right]$ with $r = \{0.02, 0.04, 0.06, 0.08\}$. Due to the similar power performance of the proposed RID test under the four different distributions, we present only the empirical power under $\mathcal{N}(0,1)$ in Figure 5. The results for the remaining three distributions are included in the supplementary material. As illustrated in Figure 5, the CQ test has the best performance, followed by our proposed RID test. Meanwhile, all tests have poor power when $r = 0.02$.

Example 3. In this example, we carry out numerical simulations to illustrate the power performance of our proposed method for the sparse setting. We use the same setup in

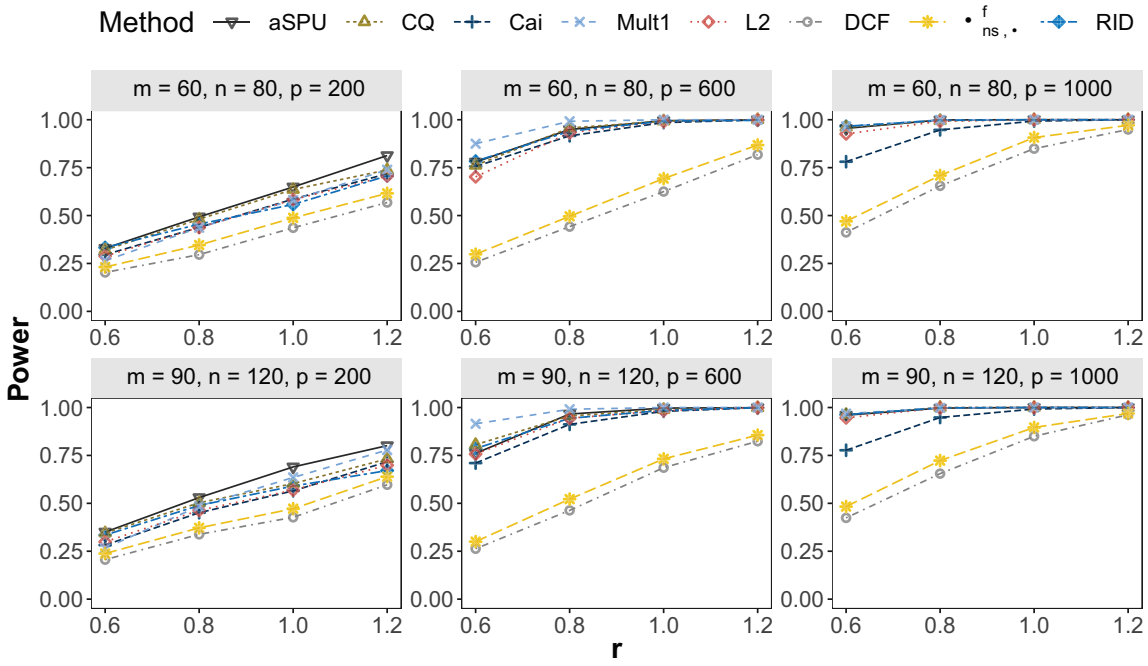


Figure 6: Empirical power when Z_{ij} follows $\mathcal{N}(0,1)$, $(m,n) = (60,80)$ and $(m,n) = (90,120)$ under different signal levels of r in **Example 3**.

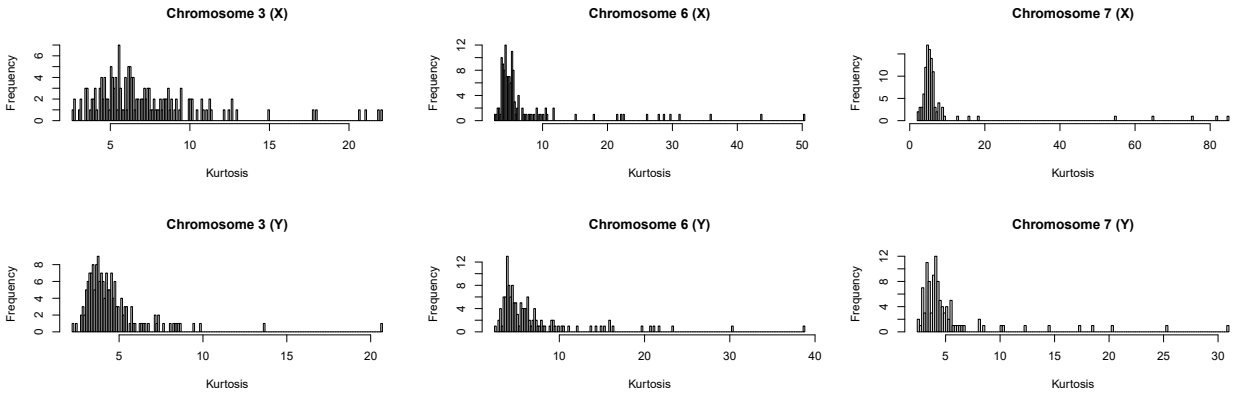
Example 1, except that we now use $r = \{0.6, 0.8, 1.0, 1.2\}$, which are considered in Chen et al. (2019), $\lfloor 0.05p \rfloor$ elements in $\boldsymbol{\mu}_2$ are set to nonzero values, which is a sparse setting and is studied in Cai et al. (2014), and $\Sigma_1 = \Sigma_2 = \Sigma$, where $\Sigma = D^{1/2}RD^{1/2}$, $R = (r_{ij})_{p\times p}$ with $r_{ij} = 1I_{i=j} + 0.5|i-j|^{-5}I_{i\neq j}$ for $1 \leq i, j \leq p$, and $D = \text{diag}(d_{11}, \cdots, d_{pp})$ with $d_{ii} \sim U(1,3)$ independently. The covariance matrice Σ structure is studied in Cai et al. (2014). Since the proposed RID test has the similar power performance under the four different distributions, we only display the empirical power under $\mathcal{N}(0,1)$ in Figure 6. For the other three distributions, the results are shown in the supplementary material. From Figure 6, we find that the power of the aSPU is the highest among all tests. Cai test, Mult1 test, and the proposed RID test have similar power when $p = 200$. Furthermore, all tests have power close to 1 when $p \geq 600$ and $r \geq 0.8$ except the DCF and $\Psi^f_{ns,\vartheta}$.

4. REAL DATA ANALYSIS

In this section, we evaluate the finite sample performance of various tests in the analysis of a breast cancer dataset from a genome-wide association study (Gravier et al., 2010), which can be downloaded from https://0-www-ncbi-nlm-nih-gov.brum.beds.ac.uk/geo/query/acc.cgi?acc=GSE19159. The dataset contains 168 pT1T2pN0 invasive ductal carcinoma patients with either good (no event five years after diagnosis: 111 patients) or poor (57 patients with early-onset metastasis) outcomes (Qiu et al., 2021). There are 2,905 genome tiling array type variables for each subject, representing the normalized log2 ratio of the Cy5 and Cy3 signals.

It has been reported that breast cancer is the most frequent malignancy in women and has been a significant source of cancer-related morbidity and mortality in women worldwide (Harbeck et al., 2019; Hendrick et al., 2019). Therefore, it is essential to identify significant genetic variants with breast cancer, and this will be helpful for the diagnosis, prevention, and treatment of breast cancer. Traditional multiple testing methods need multiple comparisons, making the genome-wide association study computationally intensive and might lead to misleading findings. Furthermore, it has been shown that there is strong evidence of polygenic effects for breast cancer (Shiovitz and Korde, 2015). Therefore, we apply various tests to analyze the genome tiling data in each of the 22 autosomes separately to better demonstrate the possible power differences between the tests. The familywise nominal significance level is set at 0.05, and we set $0.05/22 = 0.0023$ for each chromosome to consider the Bonferroni adjustment.

All of the compared methods have moment-based assumptions and according to Wang et al. (2015), they are sensitive to outliers. We refer to Wang et al. (2015); Feng et al. (2016) for specific examples demonstrating the sensitivity of such tests to the outliers. In addition, we plot the histograms of the marginal kurtosises for chromosomes 3, 6, and 7 in Figure 7, which clearly demonstrate that some gene expression levels have heavy tails due to their kurtosises being much larger than 3. Therefore, prior to performing any test, we use an



(a) (b) (c)

Figure 7: (a) Histogram of marginal kurtosises of $\mathbf{X}, \mathbf{Y}$ for 2,905 genes in chromosome 3; (b) Histogram of marginal kurtosises of $\mathbf{X}, \mathbf{Y}$ for 2,905 genes in chromosome 6; (c) Histogram of marginal kurtosises of $\mathbf{X}, \mathbf{Y}$ for 2,905 genes in chromosome 7.

outlier detection method proposed by Ro et al. (2015) to eliminate outliers. The number of samples after the outlier removal and the number of genes after the pre-processing as in Gravier et al. (2010) are listed in Table 4.

To be concise, Table 5 presents some representative results. We also plot the differences of sample means between two groups for the selected chromosomes in Figure 8. It can be seen that for chromosome 3, all methods yield p-values less than $0.05/22 = 0.0023$. For chromosome 6, RID is the only test indicating a significant signal (p-value=3.22×10^{-5}). For chromosome 7, all methods except Cai and DCF lead to significant p-values, supporting the possible lower power of Cai and DCF tests.

Having observed the significant differences, we further explore different signal structures. From Figure 8, we can see that the number of negative mean differences is comparable with those of positive mean differences on chromosome 3 while most signals are relatively strong. These signals are easily detected by most methods. In contrast, chromosome 6 has a dense-but-weak signal structure and has a large proportion of the negative mean differences (81.34%). As expected from our theoretical and simulation results, such signals

Table 4: After removing the outlier samples, the total number of samples and the number of genes after pre-processing on each Chromosome (Chr). p denotes the dimension and m, n denote the sample sizes after eliminating outliers.

Chr	p	m	n	**Chr**	p	m	n	**Chr**	p	m	n	**Chr**	p	m	n
1	374	82	42	**7**	105	68	37	**13**	84	66	34	**19**	43	81	32
2	233	69	35	**8**	112	57	37	**14**	73	70	39	**20**	88	67	30
3	166	64	39	**9**	112	67	38	**15**	77	67	30	**21**	43	69	36
4	139	74	38	**10**	132	66	35	**16**	76	72	35	**22**	68	77	41
5	185	67	34	**11**	133	65	32	**17**	156	71	33				
6	134	61	34	**12**	191	69	37	**18**	73	66	43				

are challenging to the existing methods. This application confirms that RID is particularly useful when nonzero signals are extremely dense with nearly the same sign. Also importantly, our result is consistent with the literature supporting the role of the genes on chromosome 6 in breast cancer (Noviello et al., 1996; Tao et al., 1999).

Comparing the signal structures between chromosome 6 and chromosome 7, we can see that although both have dense signals, the signal magnitude on chromosome 7 is much stronger. This indicates that the power of RID is stable when the signal is relatively weak, whereas the other methods lose power in detecting weak signals. Thus, our data analysis demonstrates that our proposed RID fills a gap in the existing methodology by introducing a test that is powerful in a setting when the existing methods are not powerful, i.e., when nonzero signals of the difference between two mean vectors are weakly dense with nearly the same sign.

Table 5: The p-values of the various methods applied to the breast cancer data.

	Method							
Chromosome	aSPU	CQ	Cai	Mult1	L2	DCF	$\Psi^{f}_{ns,\vartheta}$	RID
3	0	0	4.06×10^{-4}	0	1.41×10^{-9}	4.6×10^{-4}	0	3.55×10^{-7}
6	2.10×10^{-2}	2.91×10^{-3}	9.92×10^{-2}	3.6×10^{-3}	1.70×10^{-2}	2.15×10^{-1}	2.00×10^{-1}	8.50×10^{-5}
7	1.01×10^{-13}	3.02×10^{-8}	9.05×10^{-3}	0	2.21×10^{-4}	3.43×10^{-3}	4.00×10^{-3}	4.23×10^{-10}

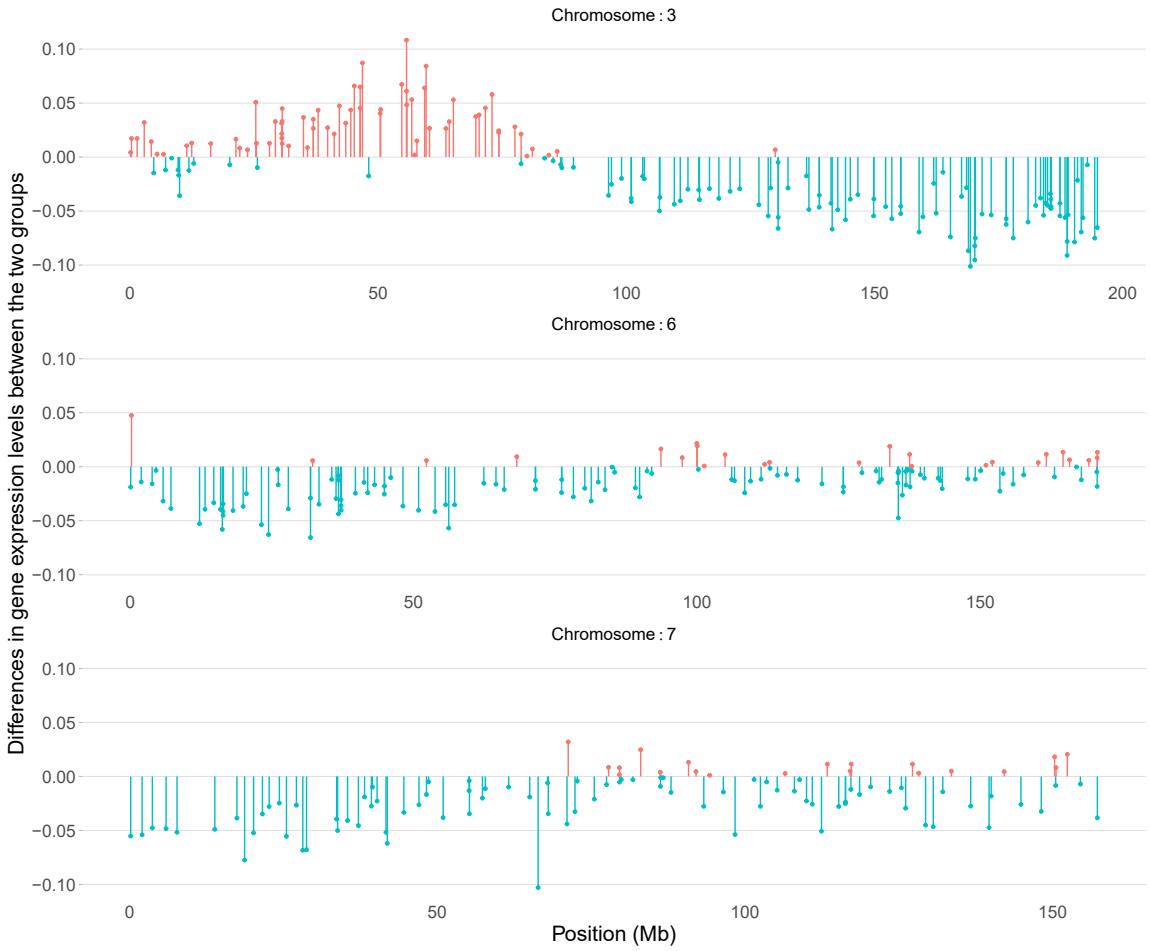


Figure 8: Differences in gene expression levels between the two groups for chromosomes 3, 6, 7. The X-axis represents the position (Mb) of the clones along the corresponding chromosome.

5. DISCUSSION

A variety of methods have been developed for testing the equality of mean vectors in two samples. As summarized in Table 1, this problem remains to be an active research topic in statistics (Chen et al., 2019; Xue and Yao, 2020; Zhang et al., 2020). Despite meaningful progress, further research is warranted. For example, there is no powerful test when nonzero signals are weakly dense with nearly the same sign or when there are more dense or only weakly dense nonzero signals (Xu et al., 2016). The first contribution of this work is to fill in this gap by developing a powerful test to detect such signals. We provide theoretical and numerical results that convincingly demonstrate that this goal is achieved. While searching for this solution, we use the random integration of the difference technique, enabling us to unify many existing methods. This is the second significant contribution of this work because this unified framework helps us understand when and why a test is powerful. By re-analysis a real dataset, we illustrate how our proposed test may discover insightful information.

It is noteworthy that there are further issues to investigate for our proposed method. Firstly, in our simulation studies and real data analysis, we use the p-dimensional independent density function with $\alpha_1 = \cdots = \alpha_p = 2p^{-3/8}$ and $\beta_i = \sqrt{2}(p+i)/p, i = 1, \cdots, p$ as the weight function. As a result, RID combines a weighted L_2-norm-based test and a burden test. Our empirical results support these choices in their effectiveness for constructing a powerful RID test. The test is still reliable when there are many small to moderate componentwise differences or when nonzero signals are weakly dense with nearly the same sign. Nonetheless, there are other choices for the weight function. An interesting future topic is to consider a p-dimensional independent density function with an adaptive choice α_i, β_i or other choices of weight function. Secondly, a weight function is given by a density function with independent components in this paper. We will consider a different measure with dependent components as further work. Finally, we investigate the fourth setting in Remark 1 in order to increase the power for extremely dense nonzero weak signals of nearly identical sign. We will consider the other settings in Remark 1 in greater detail to ensure that they meet the

requirements of practical applications.

SUPPLEMENTARY MATERIALS

We defer the technical proofs and details to the Supplementary Materials. Additional simulation results and analysis of the real dataset are also presented in the Supplementary Materials.

REFERENCES

Ayyala, D. N., Frankhouser, D. E., Ganbat, J.-O., Marcucci, G., Bundschuh, R., Yan, P., and Lin, S. (2015), "Statistical methods for detecting differentially methylated regions based on MethylCap-seq data," *Briefings in Bioinformatics*, 17(6), 926–937.

Bai, Z., and Saranadasa, H. (1996), "Effect of high dimension: by an example of a two sample problem," *Statistica Sinica*, 6, 311–329.

Bollerslev, T., Meddahi, N., and Nyawa, S. (2019), "High-dimensional multivariate realized volatility estimation," *Journal of Econometrics*, 212(1), 116–136.

Cai, T. T., Liu, W., and Xia, Y. (2014), "Two-sample test of high dimensional means under dependence," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2), 349–372.

Chang, J., Zheng, C., Zhou, W.-X., and Zhou, W. (2017), "Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity," *Biometrics*, 73(4), 1300–1310.

Chen, S. X., Li, J., and Zhong, P.-S. (2019), "Two-sample and ANOVA tests for high dimensional means," *The Annals of Statistics*, 47(3), 1443–1474.

Chen, S. X., and Qin, Y.-L. (2010), "A two-sample test for high-dimensional data with applications to gene-set testing," *The Annals of Statistics*, 38(2), 808–835.

Chen, S. X., Zhang, L. X., and Zhong, P. S. (2010), "Tests for high-dimensional covariance matrices," *Journal of the American Statistical Association*, 105(490), 810–819.

Feng, L., Zou, C., and Wang, Z. (2016), "Multivariate-sign-based high-dimensional tests for the two-sample location problem," *Journal of the American Statistical Association*, 111(514), 721–735.

Gravier, E., Pierron, G., Vincent-Salomon, A., Gruel, N., Raynal, V., Savignoni, A., De Rycke, Y., Pierga, J.-Y., Lucchesi, C., Reyal, F. et al. (2010), "A Prognostic Dna Signature for T1t2 Node-Negative Breast Cancer Patients," *Genes, Chromosomes & Cancer*, 49(12), 1125–1134.

Gregory, K. B., Carroll, R. J., Baladandayuthapani, V., and Lahiri, S. N. (2015), "A two-sample test for equality of means in high dimension," *Journal of the American Statistical Association*, 110(510), 837–849.

Hall, P. (1984), "Central limit theorem for integrated square error of multivariate nonparametric density estimators," *Journal of Multivariate Analysis*, 14(1), 1–16.

Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Ruddy, K., Tsang, J., and Cardoso, F. (2019), "Breast cancer," *Nat Rev Dis Primers*, 5(1), 66.

Hendrick, R. E., Baker, J. A., and Helvie, M. A. (2019), "Breast cancer deaths averted over 3 decades," *Cancer*, 125(9), 1482–1488.

Jiang, Y., Wen, C., Jiang, Y., Wang, X., and Zhang, H. (2022), "Use of random integration to test equality of high dimensional covariance matrices," *Statistica Sinica*, p. DOI:10.5705/ss.202020.0486.

Kim, Y., and Lesser, V. (2008), Finding Minimum Data Requirements Using Pseudo-Independence,, in *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 2, IEEE, pp. 57–64.

Lam, C., and Yao, Q. (2012), "Factor modeling for high-dimensional time series: inference for the number of factors," *The Annals of Statistics*, 40(2), 694–726.

Le Bihan, D., Mangin, J.-F., Poupon, C., Clark, C. A., Pappata, S., Molko, N., and Chabriat, H. (2001), "Diffusion tensor imaging: concepts and applications," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 13(4), 534–546.

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Team, E. L. P., Christiani, D. C., Wurfel, M. M., Lin, X. et al. (2012), "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies," *The American Journal of Human Genetics*, 91(2), 224–237.

Li, H., Aue, A., Paul, D., Peng, J., and Wang, P. (2020), "An adaptable generalization of Hotelling's T^2 test in high dimension," *Annals of Statistics*, 48(3), 1815–1847.

Li, J., and Chen, S. X. (2012), "Two sample tests for high-dimensional covariance matrices," *The Annals of Statistics*, 40(2), 908–940.

Noviello, C., Courjal, F., and Theillet, C. (1996), "Loss of heterozygosity on the long arm of chromosome 6 in breast cancer: possibly four regions of deletion.," *Clinical Cancer Research*, 2(9), 1601–1606.

Pan, W., and Shen, X. (2011), "Adaptive tests for association analysis of rare variants," *Genetic Epidemiology*, 35(5), 381–388.

Pan, W., Tian, Y., Wang, X., and Zhang, H. (2018), "Ball divergence: nonparametric two sample test," *The Annals of Statistics*, 46(3), 1109–1137.

Qiu, T., Xu, W., and Zhu, L. (2021), "Two-sample test in high dimensions through random selection," *Computational Statistics & Data Analysis*, 160, 107218.

Ro, K., Zou, C., Wang, Z., and Yin, G. (2015), "Outlier detection for high-dimensional data," *Biometrika*, 102(3), 589–599.

Shiovitz, S., and Korde, L. A. (2015), "Genetics of breast cancer: a topic in evolution," *Annals of Oncology*, 26(7), 1291–1299.

Srivastava, M. S., and Du, M. (2008), "A test for the mean vector with fewer observations than the dimension," *Journal of Multivariate Analysis*, 99(3), 386–402.

Srivastava, M. S., Katayama, S., and Kano, Y. (2013), "A two sample test in high dimensional data," *Journal of Multivariate Analysis*, 114, 349–358.

Tao, L., Elkahloun, A. G., Nguyen, H. T., Chen, Y., Lippman, M. E., and Su, Y. A. (1999), "Isolation of chromosome 6-encoded differentially expressed genes associated with breast cancer cell lines MDA-MB-231 and MDA/H6 by cDNA microarray," *Nature Genetics*, 23(3), 77–77.

Wang, L., Peng, B., and Li, R. (2015), "A high-dimensional nonparametric multivariate test for mean vector," *Journal of the American Statistical Association*, 110(512), 1658–1669.

Wang, R., and Xu, W. (2022), "An approximate randomization test for high-dimensional two-sample Behrens-Fisher problem under arbitrary covariances," *Biometrika*, p. DOI: https://doi.org/10.1093/biomet/asac014.

Wu, Y., Genton, M. G., and Stefanski, L. A. (2006), "A multivariate two-sample mean test for small sample size and missing data," *Biometrics*, 62(3), 877–885.

Xu, G., Lin, L., Wei, P., and Pan, W. (2016), "An adaptive two-sample test for high-dimensional means," *Biometrika*, 103(3), 609–624.

Xue, K., and Yao, F. (2020), "Distribution and correlation-free two-sample test of high-dimensional means," *Annals of Statistics*, 48(3), 1304–1328.

Yu, X., Li, D., Xue, L., and Li, R. (2022), "Power-enhanced simultaneous test of high-dimensional mean vectors and covariance matrices with application to gene-set testing," *Journal of the American Statistical Association*, (just-accepted), 1–39.

Zhang, J.-T., Guo, J., Zhou, B., and Cheng, M.-Y. (2020), "A Simple Two-Sample Test in High Dimensions Based on L_2-Norm," *Journal of the American Statistical Association*, 115(530), 1011–1027.